



Co-Chairperson: Jim McKenna, Port of Portland
Co-Chairperson: Bob Wyatt, NW Natural
Treasurer: Fred Wolf, Arkema

September 1, 2006

Chip Humphrey
Project Manager
U.S. Environmental Protection Agency
811 SW Sixth Avenue, 3rd Floor
Portland, OR 97204

Eric Blischke
US EPA
811 SW 6th Avenue, 3rd Floor
Portland, OR 97204

Re: Portland Harbor Superfund Site; Administrative Order on Consent for Remedial Investigation and Feasibility Study; Docket No. CERCLA-10-2001-0240. Ecological Risk Assessment Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests

Dear Chip and Eric:

Thank you for your letter of July 6, 2006 that provided comments to the Ecological Risk Assessment Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests (Benthic Interpretive Report). LWG has reviewed all the comments and is submitting a detailed response to each of the comments in the attached document.

LWG agrees with the approach for the Round 2 Comprehensive Site Summary and Data Gaps Analysis Report (Round 2 Report) to define sediment toxicity recommended by EPA and its partners. Both the floating percentile method (FPM) and the alternative set of logistic regression models (LRMs) developed by NOAA will be used as lines of evidence in assessing risks to benthic community in Portland Harbor. To avoid any potential discrepancies in predictions based on the FPM as presented in the Benthic Interpretive Report and in the requested revision of the FPM (including NJ qualified data and 1/2 DL in sums of chemicals), the FPM is currently being revised accordingly. The revised FPM will be used for the Round 2 Report. LWG is continuing to evaluate the use of TPH SQV as a surrogate for total PAHs and related issues with respect to appropriate threshold values for PAHs. LWG looks forward to further discussion of these issues with EPA and its partners.

For the baseline ecological risk assessment and remedial investigation report, LWG is proposing the following approach for integrating the results of the FPM and the LRMs to develop a predictive line of evidence for assessing risks to the benthic community in Portland Harbor. The LRMs use the *Hyalella* 28-day growth and survival endpoint and a larger national freshwater database. The LRMs will be calibrated to the Effects Level 2 for the Portland Harbor data¹. Sediment concentrations below the lowest threshold

¹ LWG is working on establishing the appropriate Pmax values for the threshold values based on Effect Level 2. Current analysis suggests Pmax values of 0.4 and 0.6 on Effects Level 2 provide the best reliability,

121 NW Everett Portland OR 97209 ♦ PO Box 3529 Portland OR 97208

USEPA SF



1482339

value will be deemed non-toxic to the benthic community based on the LRM, concentrations above the second threshold value will be deemed toxic and concentrations between the two values will be deemed indeterminate. The revised FPM uses three individual endpoints (*Chironomus* survival, *Chironomus* growth, and *Hyalella* survival) and is based on site-specific data only. Similar to the LRMs two thresholds will be used to assess risks to the benthic community. LWG is currently revising the FPM including establishing the appropriate threshold values based on a reliability analysis at Effects Level 2 and 3.

Risks to the benthic community will be assessed through the following weight-of-evidence approach:

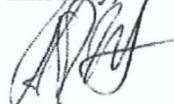
- 1) The sediment toxicity testing lines of evidence will be weighted such that they will override other lines of evidence at the locations where they were collected.
- 2) For areas where no sediment toxicity testing was performed, the Portland Harbor-specific predictive toxicity models will be weighted such that they will override other lines of evidence where the two models agree, or where one model gives a "conclusive" prediction (i.e., not toxic or toxic) and the other model gives an "inconclusive" prediction (see Table 1).

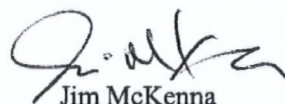
Table 1. Decision matrix for toxicity classification based on the two models

		LRM		
		Not-toxic	Indeterminate	Toxic
FPM	Toxic Classification			
	Not-toxic	Not-toxic	Not-toxic	Indeterminate
	Indeterminate	Not-toxic	Indeterminate	Toxic
	Toxic	Indeterminate	Toxic	Toxic

In areas classified as "indeterminate" by the Portland Harbor-specific predictive toxicity models (see Table 1), other lines of evidence will be used to assess benthic community risk. These include tissue concentrations data and tissue concentrations data predicted from water concentrations (with BSAFs). Other SQVs have already been assessed as a line of evidence in the Benthic Interpretive Report (Section 3.0) and rejected (i.e., assigned a weight of zero) because they were determined to be unreliable relative to the LRM and FPM models, which provide site-specific SQVs. LWG is looking forward to continued discussions with EPA and its partners about the approach for assessing risks to the benthic community in Portland Harbor.

Sincerely,


Bob Wyatt
Co-Chair


Jim McKenna
Co-Chair

cc: Jim Anderson, ODEQ (with enclosure)
Dick Pedersen, ODEQ (email only)
Mikell O'Mealy, ODEQ (email only)
Dana Davoli, USEPA (email only)
Joe Goulet, USEPA (email only)
Kristine Koch, USEPA (email only)
Rick Kepler, ODFW (email only)
Ted Buerger, USFWS (email only)

Rob Neely, NOAA (email only)
Ben Shorr, NOAA (email only)
Ron Gouguet, NOAA (email only)
Preston Sleeper, DOI (email only)
Patty Howard, CRITFC (email only)
Brian Cunninghame, Confederated Tribes of Warm Springs (email only)
Rose Longoria, Confederated Tribes of Yakama Nation (email only)
Jeff Baker, Confederated Tribes of Grand Ronde (email only)
Lisa Bluelake, Confederated Tribes of Grand Ronde (email only)
Tom Downey, Confederated Tribes of Siletz (email only)
Billy Barquin, Confederated Tribes of Siletz (email only)
Audie Huber, Confederated Tribes of Umatilla (email only)
Erin Madden, Nez Perce Tribe (email only)
Valerie Lee, Environment International (email only)
LWG Repository
LWG Legal

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
General Comments			
1	EPA would like to commend the LWG on the amount of effort that went into preparation of the Ecological Risk Assessment Interpretive Report: Estimating Risks To Benthic Organisms Using Predictive Models Based On Sediment Toxicity Tests (Benthic Interpretive Report). In general, EPA believes that LWG's proposed approach will serve as a useful tool in assessing risk and informing remedial decision making at the Portland Harbor site. EPA has developed detailed comments on the predictive models described in the report and recognizes that the project schedule does not allow time for the comments to be incorporated into the evaluation of benthic toxicity planned for the Round 2 Comprehensive Site Summary and Data Gaps Analysis Report (Round 2 Comprehensive Report). In order to avoid schedule delays, EPA recommends incorporating the results of the predictive model into the Round 2 Comprehensive Report as presented with the following modifications:	No category needed	No comment needed
2	NOAA developed alternative logistic regression models, using a larger freshwater database for the <i>Hyalella</i> 28-day growth and survival endpoint and calibrated these models to the Level 2 Effect Level in the Portland Harbor data. EPA notes that both approaches were reasonably successful at developing a predictive relationship between sediment chemistry and toxicity; the two predictive modeling approaches were in agreement approximately 75% of the time and are useful for focusing in on areas where sediment contamination is likely to pose a risk to the benthic community. EPA believes that this alternative set of logistic regression models should be applied by the LWG to the Portland Harbor data set to improve the predictive ability of these tools.	3	As stated in the proposed overarching benthic interpretive approach LWG agrees with using the alternative set of logistic regression models provided that the reliability of these models are equal to or better than the LRM version presented in the Benthic Interpretive Report (March 17, 2006).
3	The approach recommended by the LWG (Floating Percentile Method) should be applied in conjunction with the alternative logistic regression models developed by NOAA as complimentary lines of evidence. Areas where both models predict risk or do not predict risk should be identified as such. Areas where the models are not in	3	LWG agrees with using both models and the approach outlined in the EPA comment (for further details see the proposed overarching benthic interpretive approach).

LWG COMMENT CATEGORIES

1 – Strongly Disagree; cannot accept

2 – Disagree but can accept

3 – Agree

4 – Further internal discussion is needed

5 – Unclear; requires clarification from EPA

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
	agreement should be identified as areas of indeterminate risk. Areas of indeterminate risk should be refined based on other lines of evidence used to evaluate risk to the benthic community.		
4	The approach recommended by the LWG includes a proposed sediment quality values (SQV) of 1,270 mg/kg for total PAHs. This concentration is more than 50 times the concentration of the consensus based probable effects concentration (PEC) of 23 mg/kg developed by MacDonald and Ingersoll. As a result this value should not be applied to the data set. The LWG recommended floating percentile method should rely on the SQV developed for diesel range hydrocarbons as a surrogate for total PAHs.	4	LWG is continuing to evaluate the use of TPH SQV as a surrogate for total PAHs
5	The Round 2 Report should use the floating percentile methodology and the refined logistic regression methodology to identify areas of potential concern based on risks to the benthic community. Refinements to the predictive approach outlined in the attached comments should be used in conjunction with the results of the Round 2 report to identify additional data needs that will improve the models' ability to predict risks to the benthic community. These data gaps should be filled as part of the Round 3B sampling effort to be completed in 2007. EPA comments on the predictive models should be incorporated into the next iteration of the Benthic Toxicity Interpretive Report to be presented in the baseline ecological risk assessment and remedial investigation report.	3	LWG agrees with using the FPM and the refined LRM to predict sediment toxicity at stations where toxicity test data are not available to support the benthic toxicity assessment in the Round 2 Report. As presented in the overarching benthic interpretive approach the FPM is currently being revised to include NJ qualified data and ½ DL in sums of chemicals. The two models will be used for the baseline ecological risk assessment and remedial investigation report.
6	Focus the Modeling Efforts: This report recommends focusing on the floating percentile method for future modeling efforts. As described above, the LWG and NOAA models are in agreement approximately 75% of the time. As a result, EPA believes that both models should be utilized as complimentary lines of evidence. Areas where both models predict risk or do not predict risk should be identified as such. Areas where the models are not in agreement should be identified as areas of indeterminate risk. Areas of indeterminate risk should be refined based on other lines of evidence including empirical estimates of benthic toxicity using bioassays; comparison of benthic tissue data (empirical measurements or modeled through application of BSAFs) to	3	LWG agrees with using the revised LRM and the FPM as lines of evidence in assessing risks to the benthic community. LWG also agrees on using other site-specific data such as benthic tissue data as other lines of evidence. However, LWG have assessed other SQGs from the literature (see Section 3.0 in Benthic Interpretive Report) as lines of evidence and rejected them (assigned a weight of zero) because they were assessed as unreliable relative to the FPM and LRM models.

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT NO.	EPA COMMENT	COMMENT CATEGORY	NOTES
	tissue TRVs; comparison to consensus, empirical and/or empirical based sediment quality guidelines (SQGs); and comparison of transition zone water data (empirical measurements or modeled through application of partitioning equations) to AWQC or literature values.		
7	<i>Hyalella</i> growth and survival endpoint: The Lower Willamette Group (LWG) proposes to disregard the results of the <i>Hyalella</i> growth and survival (pooled) endpoint. LWG supports this proposal based on "difference from other endpoints" and "no correlation with mortality endpoint." Yet these are precisely the reason that multiple test endpoints are required (because different test endpoints may show different sensitivities to different chemical mixtures). However, there was substantial agreement between the <i>Hyalella</i> and <i>Chironomus</i> pooled endpoints for samples that showed an extreme degree of toxicity (e.g., < 50% of control) in either test. The "lack of correlation to Chemicals of Concern" and the "effect of percent fines" may be more related to the different contaminant mixtures and gradients in the Portland Harbor study area. In a complex environment with multiple chemical mixtures and gradients with limited numbers of samples from any one area, a lack of correlation between a test endpoint and individual chemicals does not necessarily imply that toxicity is not related to chemical contamination. This is supported by the differences in chemicals that "set" the different models for the same sample (for example, the chemical with highest ratio of concentration to floating point value for a sample may be a phthalate, while the chemical with the highest probability of toxicity in logistic regression models may be ammonia or DDT for the <i>Hyalella</i> pooled model or PCBs or cadmium for the <i>Chironomus</i> pooled model). Because each contaminant can be considered as an indicator of toxicity for the chemical mixtures, it is not surprising that generic indicators such as percent fines, ammonia, or sulfides are good predictors of toxicity.	1	As stated in the overarching benthic interpretive approach, LWG proposes using the revised LRM based on the <i>Hyalella</i> pooled endpoint and the FPM based on <i>Chironomus</i> growth, <i>Chironomus</i> mortality, and <i>Hyalella</i> mortality endpoints as separate lines of evidence in assessing risks to the benthic community. By using both models all information from the two bioassays will be utilized. LWG disagrees with including the <i>Hyalella</i> growth endpoint in the FPM because it degrades the performance of the model and its reliability in predicting toxicity.
8	Proposed total PAH threshold values: The proposed Effects Level 2 and Effects Level 3 concentrations for total PAH, which represent AET values, are unreasonably high (1270 ppm DW) and significantly higher	4	LWG is continuing to evaluate the use of TPH SQV as a surrogate for total PAHs, and the identification of appropriate PAH threshold values.

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
	than other published values. For example, the proposed value exceeds the consensus-based freshwater PEC for Total PAH (22.8 ppm DW; MacDonald et al 2000) by more than a factor of 50. Of the samples exceeding the PEC value, 73% have a Level 2 response or greater in one or both of the pooled endpoints and 86% for samples with at least 25% fines. If we exclude the <i>Hyalella</i> growth endpoint, 62% of the samples exceeding the PEC have Level 2 or greater response compared to 65% of the samples with diesel concentrations exceeding the proposed FPM value of 340 ppm. While diesel concentrations may be a slightly better predictor of toxicity than total PAH for this dataset, total PAH concentrations much lower than the proposed AET values are reliable predictors of toxicity. The proposed values for total PAH serve no useful purpose and should be discarded.		
9	Inclusion of Appropriate Data in the Model: Data for which bioavailability is an issue should not be included in the predictive model. For example, high concentrations of PAHs may be detected in the sediments, but are bound up in a less bioavailable fraction such as pencil pitch. This issue was raised previously by EPA and its partners in the context of including Port of Portland Terminal 4 data in the analysis for this reason. Including these samples in the analysis can greatly skew the model results, because effect is not correlated with bioavailable fractions in the sediment. Based on our review of the report, the inclusion of GASCO effect / concentration data may skew the model results. This site has the potential to contain many different bound PAH contamination including pencil pitch. However, these samples were still included in the model analysis. This results in the inclusion of "no hits" with very high concentrations of PAHs. Looking at the highest no-hit concentrations, the top 6 samples are all in the vicinity of the GASCO site. Examples include G-264, 1,708,600 ppb total PAHs, G-301, 1,250,500 ppb, and G178, 470,060 ppb. Conditions off GASCO are confounded by the mixture in sediments of these less-bioavailable fractions such as pencil pitch and weathered tar pieces tar along with more fresh PAH and coal tar fractions that are	No category needed	This comment has been withdrawn (personal communication Lisa Saban and Eric Blichkle).

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
	more bioavailable and elicit effects. These two conditions may be teased out by a re-analysis of the sediment samples off the site. Conditions off GASCO can also lead to variance in the toxicity test results that are too high to detect anything but very large differences (low power), resulting in statistically indeterminate results. The GASCO site had the highest incidence of indeterminate samples at all effects levels (Figure 2-2). If these effects cannot be teased apart, we could simply omit samples off the GASCO site from the analysis. For this site, it may be clear that due to the variability in the forms of the contamination that we cannot accurately predict toxicity off this facility.		
10	Three Tiered Framework: Based on the inherent reliability problems associated with development of a single SQV, EPA recommends calculating two screening values; a low screen below which a sample shouldn't be toxic and a high screen above which it should be toxic. Optimization should be possible at these two ends of the spectrum. As noted previously, the two models are generally in agreement in predicting very toxic samples and those that are clearly non-toxic samples. However, we don't agree on the classification of the samples that fall in between these classifications. The values that fall in between these two classifications would be classified as "indeterminate", and would require empirical toxicity testing or the use of additional lines of evidence. The LRM is well suited to this. It could also be done with the FPM (as was done for the DRAFT Washington Freshwater criteria).	3	LWG agrees with using two screening values for each of the two models. In the LRM approach, two points will be selected along a single curve to generate its two screening levels, while the FPM develops two separate models for the two levels. The two proposed screening values will be based on the effects levels 2 and 3. For further details see the proposed overarching benthic interpretive approach.
11	Alternative Approaches For Subsets of PH Sediments: As stated in the March 18 th work plan (Section 9.2), there are areas for which the predictive approach would not apply in Portland Harbor. This could include the physical form of the contaminant (as mentioned above), or the localized presence of contaminations over smaller spatial scales in the ISA (e.g. pesticides around RM 7). The work plan states models or other approaches would be developed for these areas. However, this was not included in the report. Also, areas where volatile chemicals were detected in sediments and may be contributing to toxicity, but not evaluated in this report should be examined.	3	As presented in the proposed overarching benthic interpretive approach, the two models (FPM and revised LRM) will be used as two components of a single line of evidence (i.e., predictive approach) in the weight of evidence approach for the benthic toxicity assessment. For areas where the results of the two models conflict and yield an indeterminate overall result (see matrix in discussion of overarching benthic interpretive approach), other lines of evidence such as the benthic tissue data will be used in assessing risks to the benthic community. Performance of additional sediment toxicity tests in these areas may be needed to adequately assess the risks

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
			to the benthic community.
12	Level 1 Biological Effects Level: The report states " <i>it is recommended that Level 1 not be used to set SQVs for Portland Harbor because it is relatively unreliable in accurately predicting effects and well below the cleanup levels set at other regional Superfund sites.</i> " EPA agrees that Level 1 Biological Effects Level values should not be used as target cleanup levels. However, Level 1 values should not be discarded, as they represent concentrations associated with low level effects and provide useful information for defining areas of concern. The incidence of Level 1 or greater effects increases with increasing probability of toxicity.	1	The screening values for the two models will be based on effects similar to the effects levels 2 and 3 in the Benthic Interpretive Report (March 17, 2006). Because the sediment toxicity tests used in Portland Harbor were not designed to detect less than about 20% difference from control, the effect level 1 which allows for only a 10% difference between site sample and control responses should not be used to define areas of concern.
13	Single-threshold evaluation of reliability: The report relies exclusively on a single-threshold evaluation of "reliability" of sediment quality guidelines. The conceptual model that a single value can accurately distinguish between "good" and "bad" samples, while perhaps desirable, is not consistent with most environmental data. EPA agrees that minimizing false negatives and false positives is an important goal, but concentration-response relationships are usually continuous and multiple thresholds may provide better separation of false positive and negative concentrations. For continuous models, such as the logistic regression model, an evaluation based on a single-threshold loses important information.	5	Two screening values will be proposed for both the revised LRM and the revised FPM after completion of the analyses. The reliability assessment of the two models in the Benthic Interpretive Report included all available reliability measures.
14	LRM model development: The logistic regression models were developed following the published approach developed by NOAA and EPA (Field et al. 1999; Field et al 2002; EPA 2005). The model development presented in the report did not address exclusion of chemical models that resulted in a high degree of false positives or adjustments to the screening approach to reduce the influence of a small number of non-toxic samples with very high chemical concentrations, which was particularly problematic for PAHs. The models were evaluated for reliability using the single threshold approach. Although this evaluation provides some useful information, reducing the evaluation to a single threshold does not take full	3	The issues raised in this comment are addressed in the proposed overarching benthic interpretive approach where LWG agrees to use the revised LRM and two screening values for each model which addresses this comment. The reliability assessment of the models will include all available reliability measures. The reliability of the revised LRM should be equal to or better than the LRM version presented in the Benthic Interpretive Report (March 17, 2006).

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT NO.	EPA COMMENT	COMMENT CATEGORY	NOTES
	advantage of the continuous concentration-response relationship.		
15	NOAA developed alternative logistic regression models, using a larger freshwater database for the <i>Hyalella</i> 28-day growth and survival endpoint and calibrated these models to the Level 2 Effect Level in the Portland Harbor data. EPA believes that this alternative logistic regression model should be applied by the LWG to the Portland Harbor data set to improve the predictive ability of these tools.	3	LWG agrees to use the revised LRM if the reliability of this version is equal to or better than the LRM version presented in the Benthic Interpretive Report (March 17, 2006).
16	Recommended FPM values: The recommended FPM values are based on 3 individual endpoints (<i>Chironomus</i> survival, <i>Chironomus</i> growth, and <i>Hyalella</i> survival), excluding results for the <i>Hyalella</i> growth endpoint and for the combined (pooled) growth and survival endpoints for both test species. The pooled results are important to consider, because growth and survival are not independent measures. (See previous discussion of the rationale for including the <i>Hyalella</i> growth and survival combined endpoint.)	1	As presented in the overarching benthic interpretive approach LWG proposes using the FPM based on <i>Chironomus</i> growth, <i>Chironomus</i> mortality, and <i>Hyalella</i> mortality endpoints and the revised LRM based on the <i>Hyalella</i> pooled endpoint two components of a one line of evidence in assessing risks to the benthic community. By using both models all information from the two bioassays will be utilized. LWG disagrees with including the <i>Hyalella</i> growth endpoint in the FPM because it degrades the performance of the model and its reliability.
17	Several of the recommended FPM values have the same concentration for Level 2 and Level 3 Effects. This indicates that these values are at the upper end of the concentration-response relationship and thus may be considered extreme effect concentrations.	5	By definition, Level 2 and Level 3 responses are within the lower half of the dose-response curve, as all of these effects are less than 30%. What happens is that there tend to be multiple thresholds within the data distribution – if the false negative rate is graphed with increasing concentration, it is not a continuous curve but more a series of jumps. When the model reaches one of these plateaus it tends to remain there until the target false negative rate is significantly increased. Frequently, the threshold is the same for Level II effects as for Level III effects – these levels are not that different on the dose-response curve and the hit/no-hit distributions are similar. These thresholds may occur at any effects level – very low, intermediate, or high. For any value listed as a Level II or Level III SQV however, the threshold in question is by definition somewhere in the 20%-30% range of effects.
18	PEC-quotient approach: The report did not evaluate the PEC-quotient (PEC-q) approach (Ingersoll et al 2001) – one of the major approaches to developing freshwater guidelines – which has been	1	Two variations of the PEL-Q developed by Dr. Ingersoll et al. were evaluated (Appendix A). The PEL-Q would be expected to have more validity than a PEC-Q, as the PECs themselves were not

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
	applied effectively in other Superfund remedial investigations (e.g., Calcasieu Estuary, Louisiana). A quick review of the data indicate that samples with mean PEC-q's greater than 1 show a Level 1 response or greater in at least one toxicity test endpoint in 87% of the samples and at least a Level 2 response in 77% of the samples. This suggests that the PEC-q approach may be useful in contributing to the identification of areas of concern. Evaluation of the Ingersoll PEC-q should be performed to determine if it is useful for the Portland Harbor remedial investigation.		developed in a mathematically rigorous way. The PEL-Qs have also been adopted for use in at least one regulatory program. In addition to the PEL-Q evaluations, the ERM-Q was evaluated at EPA's request. All of these methods perform similarly, and the PEC-Q would not be expected to be substantially different. Each is better than the off-the-shelf SQV sets, but not as good as the site-specific models. This may be because any quotient method is more reliable if it is calibrated to the site.
19	Data Gaps: TPH was found to have good potential relationships with toxicity. However, because TPH was only analyzed at a limited number of stations, the model cannot assess this relationship (see page 21). As a result, additional TPH data may be required.	4	LWG is continuing to evaluate the use of TPH SQV as a surrogate for total PAHs
SPECIFIC COMMENTS:			
20	Page 1, Section 1.0, Introduction: There are statements made here that state that the sediment toxicity testing and derivation of sediment quality values (SQVs) form the primary lines of evidence for the benthic community, and that other lines of evidence such as tissue residue concentrations and comparison to surface water and transition zone water concentrations would be secondary lines of evidence. This text should be revised, as the weights of different lines of evidence will be developed through the development of the weighting matrix.	3	Please refer to the overarching approach proposed by LWG in the cover letter.
21	Page 5, Section 2.0, Data Quality and Organization: The report states that "petroleum data for 203 stations" were available. How were the 146 stations with matching toxicity data for petroleum analysis selected?	No category needed	The stations were selected based on proximity to potential sources (i.e., near fuel facilities) as stated in the Round 2 QAPP (June 24, 2004).
22	Page 5, Section 2.1.2, Biological Effects Definitions: The report states that "The biological effects levels used in the analyses are intended to correspond conceptually to "no effects level" (Level 1), "minor effects level" (Level 2), and "moderate effects level" (Level 3). As requested by EPA (EPA 2005a), the three levels were set at 90, 80, and 70% of	1	The toxicity tests used at the Portland Harbor were not designed to reliably detect a 10% difference from control and therefore it would be inappropriate to consider the Level 1 response as an effect level for final site-specific SQVs..

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
	<i>the response observed in the control sediment, respectively.</i> The biological effect levels are mischaracterized. A more appropriate characterization would be "minor effects level" (Level 1), "moderate effects level" (Level 2), and "severe effects level" (Level 3).		
23	Pages 5-6, Section 2.1.2, Biological Effects Definitions: Previous comments submitted by EPA have expressed concern about the selected alpha level for determining statistical significance. According to the work plan proceeding this report (<i>Estimating Risks to Benthic Organisms Using Sediment Toxicity Tests</i> , FINAL, dated March 18, 2005), an alpha level of 0.1 was to be used where it is found that test power of the dataset is low, according to ASTM guidelines (2003). Only an alpha level of 0.05 was used here. Since power is directly related to variance in the sample data, the variance in the analysis should be clearly reported and understood. To address concerns about the appropriateness of the statistical analysis to determine hits and no hits, it is recommended that the methodology outlined by Thursby et al., 1997 and Phillips et al., 2001 be followed. This approach more directly deals with issues that hinder appropriate statistical comparisons to determine statistical difference. This protocol considers performance over a large number of comparisons. MSD values are calculated to determine a critical threshold for statistically significant sample toxicity. Significant toxicity threshold values (as a percentage of laboratory control values) are presented for each species and endpoint based on the data.	3	<p>The workplan did indicate that the alpha level would be increased to 0.10 if power was found to be low at alpha = 0.05. However, changing the alpha level for a subset of the samples would lead to some minor inconsistencies within the data set, and complicate the interpretation of toxicity results because of differing statistical error rates. Following discussions with NOAA and EPA, we chose to deal with the issue of statistical power in a slightly different manner.</p> <p>The 3 effects levels are defined by statistical significance from control AND a minimum magnitude difference relative to control (Table 2-1, page 6). Non-significant results must have had sufficient power to detect a difference equivalent to the threshold used for each effect level. Specifically, for each effect level, we categorized samples as:</p> <ul style="list-style-type: none"> • Hit = statistically different from control AND test response relative to control exceeds the threshold; • No-Hit = test response relative to control is less than the threshold OR is not statistically different from control and there was sufficient power to detect the desired magnitude difference (i.e., the MDD (with beta = 0.2 and alpha = 0.05) was equal to or less than the threshold). • Inconclusive = test response relative to control is greater than the threshold AND non-significant statistical results (alpha = 0.05) AND MDD (beta = 0.2 and alpha = 0.05) was greater than the threshold. <p>There were no inconclusive samples at Levels 2 or 3, so the issue of insufficient power is only relevant for the Level 1 definition.</p> <p>The approach we used is very similar to that presented in Thursby</p>

LWG COMMENT CATEGORIES

1 – Strongly Disagree; cannot accept

2 – Disagree but can accept

3 – Agree

4 – Further internal discussion is needed

5 – Unclear; requires clarification from EPA

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
			<p>et al. (1997), which was to select a threshold considered to be important (Thursby chose 80% of the 90% minimum control survival); and compare the MSD for each test to that threshold.</p> <p>The three threshold levels used were considered to be biologically meaningful and were recommended for use by EPA (EPA memo to LWG dated 10/26/2005). Phillips et al. (2001) do not present thresholds for our test species. They likely did not intend for a project-specific 90th percentile MSD to be derived and applied to the same project, as this would result in 10% of the samples failing the criterion. What we can do is compare our thresholds (90%, 80%, and 70% relative to control) to the "statistically attainable" thresholds derived as the 90th percentile of all MSDs for each endpoint (after Phillips et al. 2001). LWG is currently working on this issue and will present the results after the analysis has been completed.</p>
24	Data should be reported as indicated in Table 2 of Phillips et al, 2001, which clearly shows the sample and control response, the sample response as a % of the control, MSD threshold, significance of t-test, and whether it was identified as toxic, non-toxic or indeterminate. This will improve the transparency of the statistical analysis, and will address several concerns associated with interpreting toxicity test data. These include:	3	LWG is currently working on station-specific toxicity test results modeled after Table 2 of Phillips et al., using the individual thresholds identified for our three Effects Levels. The results will be presented after the analysis has been completed.
25	The identification of small differences that are statistical different from the control, which may increase the probability of making a type I error (identifying a sample as toxic when in reality it is not). This reporting should eliminate cases where statistical significance is assigned in individual cases because the among-replicate variability is small in a more transparent fashion. It will allow for a better understanding for where and how much this occurred.	3	This issue (statistical significance of small differences due to small variance) was addressed with our original approach requiring a minimum threshold difference. This will be clarified with completion of the analysis requested in comment #24.
26	Samples with large variance in the data (e.g. variance lies outside the 10 th and 90 th percentiles) should be reported. Declaring a sample non-toxic in this case would lead to a greater probability of making a Type II error (saying it is non-toxic when in reality toxicity exists). This	3	This issue (lack of statistical significance of even large differences due to large variance) was addressed with our original approach requiring a minimum MDD. This will be clarified with completion of

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
	will help in understanding where areas of large variance occurred (e.g. where differences from the control exceed 20 to 25%), and further action in those areas such as re-testing.		the analysis requested in comment #24.
27	<p>This method more accurately describes Beta error through a graphic representation of the statistical power (1-B). For example, power curves can be developed using the 10th, 25th, 50th, 75th, and 90th percentiles of the variance. Power curves can be superimposed with curves showing the probability of statistical difference created from the cumulative frequency of calculated MSDs.</p> <p>This approach should explicitly define MSD for the project in a non-arbitrary manner. A better understanding of the power curves relative to the data's variance aids in decisions regarding what difference from control is appropriate for determining statistical significance. Once a threshold for significance is determined, all of the test's data is included in the acceptability analysis for that test.</p>	3/1	<p>We do not intend to change the power threshold as this would potentially require re-doing all modeling based on the hit/no-hit designations. Defining an MSD based on project specific statistical significance seems less desirable than an independent biological definition of a meaningful difference. We have compared the threshold levels recommended by EPA and used in our assessment to the distribution of MSDs (after Phillips et al 2001) to see how attainable they are given the variability of the test. The fact that there a fair number of Indeterminate samples at Level 1 and none at Level 2 indicates that a statistically attainable difference lies somewhere between 10 and 20% difference from control. LWG is currently working on additional information regarding the statistical power for these data and the results will be presented when the analysis is completed.</p>
28	<p>Page 6, Section 2.1.2, Statistical Difference Determinations: What analysis was used to determine statistical significance? The footnotes on Table 2-1 state the means of untransformed mortality or weight data was used in the definitions of effect levels. Were test and reference stations tested for normality? Were t-tests used?</p>	3	<p>The Benthic Interpretive Report will not be revised and re-submitted because of schedule limitations (not enough time before the Comp 2 report). Instead the clarification has been included here in the response to comment document.</p> <p>The untransformed data were used only in determination of whether the data met the threshold.</p> <p>Statistical tests were performed using SEDQUAL. Statistical significance was based on a 2-sample comparison between test and negative control. Residuals were evaluated for normality using Shapiro-Wilk's test, and transformed if necessary (ASIN(sqrt(x)) for mortality data; log10(x) for growth data). If transformed data failed the test for normality, a non-parametric t-test (Mann-Whitney) was used. For parametric tests, either a standard t-test or Welch's approximate t-test with separate variances was used depending on the outcome of Levene's test for equality of variances. If one sample had zero variance, then a one-sample t-test was used to</p>

LWG COMMENT CATEGORIES

1 – Strongly Disagree; cannot accept

2 – Disagree but can accept

3 – Agree

4 – Further internal discussion is needed

5 – Unclear; requires clarification from EPA

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
			compare the sample with variance to the mean of the other sample.
29	<p>Page 6, Section 2.1.2, Indeterminate Stations: EPA understands that there may be situations where low power is a problem because the variance may be too high in the test replicates to detect anything but vary large differences. Since the test responses were compared to control responses for the statistical evaluation, it is likely large variability in response came from the test sediment.</p> <p>The source of the variance should be reported here, because it could be do to variations in bioavailability related to chemical form in the environment, or due to poor sediment homogenization prior to testing.</p>	3/1	<p>The revised presentation of toxicity test results (after Table 2 in Phillips et al. 2001) will show whether the test or reference sample had higher variance. However, we do not have any information to indicate the <u>source</u> of the variance, only which sample it shows up in.</p>
30	<p>Page 6, Section 2.1.2, Biological Effects Definitions, Statistical Difference from Negative Control: For the floating percentile analysis, it would be still important to include a Level I effects level based on a statistical difference from negative control. Again, this may be more important for the floating percentile analysis (and AET derivation), especially since it is so reliant on the how we define no-hits, as apposed to hits (see page 7, second paragraph). Very small magnitude differences at the low end of the effects range may be very important for the development of the floating percentile model. The logistic regression model is not as sensitive to the omission of hits at the low range because it is the prevalence of toxic samples that primarily drive the curves, and the development of model relationships are not adversely affected by low power samples (Jay, correct me if I am wrong).</p>	1	<p>The FPM is not particularly reliant on the no-effects distribution vs. the hit distribution. Both are used equally in developing the model results. This is a fundamental difference from the AETs.</p> <p>The omission of the Level 1 effects level does not affect the results of the Level II or Level III analysis, as they are completely separate model runs (unlike the LRM).</p> <p>Level 1 as an effects level was omitted based on the low reliability at this level seen in both models, most likely due to natural and laboratory variability falling within this very low range of effects. In addition, the toxicity tests used at the Portland Harbor were not designed to reliably detect a 10% difference from control and therefore it would be inappropriate to consider the Level 1 response as an effect level.</p>
31	<p>Page 7, Section 2.1.3, Use of Historical Toxicity Data: The objectives of the modeling effort are not just to improve model reliability as defined in the footnote on page 6 (correct predictions / total stations). The results of combining historical or regional data should be presented in how it changes the endpoints the government team are interested in optimizing; including % Predicted No Hit Efficiency.</p>	1	<p>As presented in the overarching benthic interpretive approach LWG proposes using the revised LRM which includes other national historical data and the FPM which includes the site-specific data. The decision not to include regional historical data in the FPM was based on a variety of reasons. There are no chronic data in the historical site-specific data, only data for <i>Chironomus tentans</i> 10-day test is common to both. Also many of the regional historical</p>

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT NO.	EPA COMMENT	COMMENT CATEGORY	NOTES
			data sets have very incomplete chemical suites, complicating their use. At one point in the evaluation process the limited historical site-specific data was combined with the current data in the FPM (using methods and endpoints that were later modified). None of the reliability measures improved when the historical data were added to the current data.
32	<p>Page 8, Section 2.2.1 – Data Quality: The report states that “The exclusion of data with the N-qualifier primarily affected the pesticide data. Between 23 and 53% of the data for the following pesticides were excluded: aldrin, hexachlorocyclohexane (alpha-, beta-, and delta-), nonachlor (cis- and trans-), dieldrin, and methoxychlor. Between 35 and 67% of the summed data of DDD, DDE, DDT, total DDT, total chlordane, and total endosulfan were excluded.”</p> <p>Considering that some of these contaminants are known to be of importance in the Lower Willamette, further evaluation of the exclusion of the aforementioned results should be performed. For example, what percentage of the excluded data had concentrations that exceeded the 25th percentile of the detected/included data? Would including these data affect the results?</p>	3	<p>After further evaluation, it has been determined that reintroduction of the NJ data does affect the results for some chemicals (e.g., alpha- and delta-HCCH, endrin ketone, dieldrin). These NJ data will be added to the model and the model will be rerun. One chemical, methoxychlor, is shown to be non-significant once the NJ data are added. This chemical will be removed from the model runs.</p> <p>Additionally, this reanalysis indicated that DDD, DDE, and DDT have different relationships to the toxicity data, and should be included as separate variables in the model, rather than summing them all into Total DDTs. The remaining pesticide sums are still appropriate to use (e.g., total Chlordanes, total Endosulfans).</p>
33	<p>Page 8, Section 2.2.1, Data Quality: The text states that results with qualifier definitions listed in Table 2-3 were excluded. It looks like excluding samples with the “N” qualifiers excluded a lot of data (esp. pesticides). It should be confirmed that all PCB / DDT interferences in this dataset were properly re-analyzed according to previous EPA direction and the memo entitled “EPA Region 10 Guidance for Data Deliverables from Laboratories Utilizing SW-846 Methods 8081 and 8082 from the Analyses of Pesticides and PCB Aroclors”. High detection limits, or elimination of “N” qualifiers that may represent interference problems can have a significant affect on the appropriateness of any model that attempts to correlate effects with sediment concentrations. This is particularly worrisome because the text states that this exclusion primarily affected the pesticide data, and that “between 35 and 67% of the summed data of DDD, DDE, DDT, total DDT, total chlordane, and total endosulfan were excluded” (in</p>	3	<p>Because of the interferences indicated by the NJ qualifier, the addition of these data within the model may result in screening levels that have less certainty than screening levels for other chemicals. As noted above, these data will be included in the model, and this additional uncertainty will be noted in the text.</p>

LWG COMMENT CATEGORIES

- | | |
|--------------------------------------|--|
| 1 – Strongly Disagree; cannot accept | 4 – Further internal discussion is needed |
| 2 – Disagree but can accept | 5 – Unclear; requires clarification from EPA |
| 3 – Agree | |

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.

Response to EPA Comments Dated 7/6/06 on Benthic Interpretive Report

COMMENT No.	EPA COMMENT	COMMENT CATEGORY	NOTES
	addition to between 23 to 53% aldrin, hexachlorocyclohexane, nonachlor, dieldrin, and methoxychlor).		
34	Since the "N" qualifier is not an undetected value, it is unclear if this was an appropriate exclusion. It is also unclear why "N" qualifiers combined with "T" values were excluded. Is this the result of combining the results of two different analyses on one sample, where both of them were an "N"? Or, was one sample an "N" and the other a "J"? J values certainly shouldn't be excluded, so if they were combined with an "N" as a result of another analysis method shouldn't the "J" estimate take priority? Generally, EPA recommends including the N, NJ, and NJT values for modeling purposes.	3	See responses to Comments 32 and 33.
35	Page 9, Section 2.2.2, Data Organization and Reduction: The report states that " <i>The presence of non-toxic, naturally occurring crustal elements such as aluminum and selenium can confound the development of meaningful SQVs for the remainder of the analytes.</i> " It is not clear why this should be the case. This may be an issue for FPM development, but LRMs are developed independently for each chemical and the crustal elements can be included or not in the development of the maximum probability model.	3	The two models are different in that the FPM is based on mixtures of chemicals whereas LRM develops independent models for each chemical. However, when the LRM was performed as a composite model these crustal elements were excluded as in the FPM.
36	Page 9, Section 2.2.2, Data Organization and Reduction: For the FPM, aluminum and selenium should be added back into the model. The analysis shows that there is an association between aluminum and effects. I also wouldn't say that just because they are crustal elements that they are non-toxic, or that they cannot also be elevated anthropogenically. The ANOVA results for these chemicals need to be included in Table 5-2. If they are not associated with toxicity, they will drop out if appropriate.	1	As presented in the overarching benthic interpretive approach both the FPM and the revised LRM will be used to develop a predictive line of evidence for assessing risks to the benthic community in Portland Harbor. Hence, any potential toxicity associated with crustal elements will be evaluated using the revised LRM.
37	Page 11, Section 2.3.3 and Table 2-4: For some chemicals where there were elevated detection limits, the exclusion of these chemicals in contributing to the sum could underestimate the total concentration. In general, when summing chemicals ½ the detection limit should be used for non-detected values. In addition, the report states that " <i>Individual dioxins and furans (replaced by TEQ).</i> " TEQs are based	3/1	To be consistent with the risk assessments, the models are being rerun using ½ the detection limit where non-detected values are summed. LWG disagree with EPA's statement about TEQs. TEQs are based on TEFs which are a relative measure of toxicity and are

LWG COMMENT CATEGORIES

1 – Strongly Disagree; cannot accept

2 – Disagree but can accept

3 – Agree

4 – Further internal discussion is needed

5 – Unclear; requires clarification from EPA

DO NOT QUOTE OR CITE

This document is currently under review by US EPA and its federal, state, and tribal partners, and is subject to change in whole or in part.